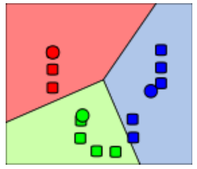
From biomolecular data to information

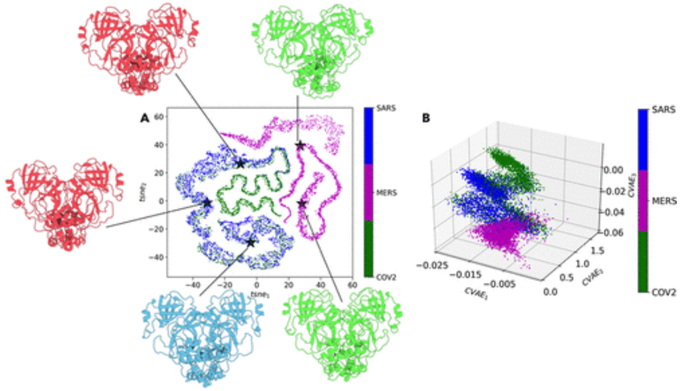
CCP5 Summer School @ University of Durham 26-27 July 2022

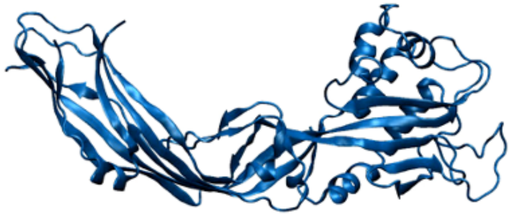
Micaela Matta
micaela.matta@kcl.ac.uk
@micaelamatta

Antonia Mey
antonia.mey@ed.ac.uk
@ppxasjsm

Matteo Degiacomi
matteo.t.degiacomi@dur.ac.uk
@MatteoDegiacomi







Schedule

Morning

09:00-11:00	Dimensionality Reduction theory and toy examples (TM)
11:00-11:30	☕ break ☕
11:30-12:30	ML Dimensionality Reduction application to protein simulations (MD)

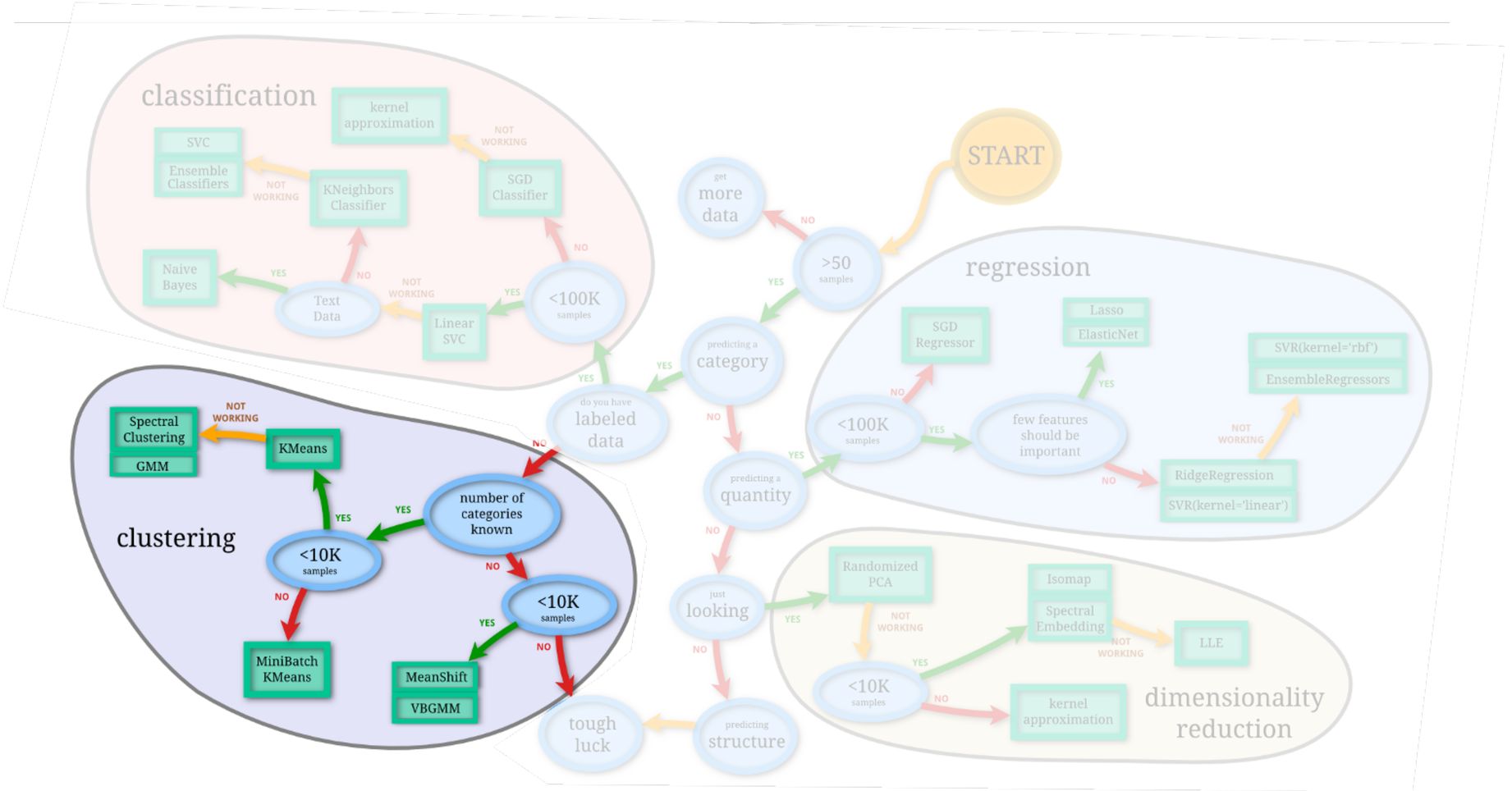
Afternoon

14:00-14:30	Clustering Theory (MD)
14:30 - 15:30	Clustering in practice (TM)
15:30 - 16:00	☕ break ☕
16:00 - 17:00	Classification problems (MD)

TM — Toni Mey

MD — Matteo Degiacomi

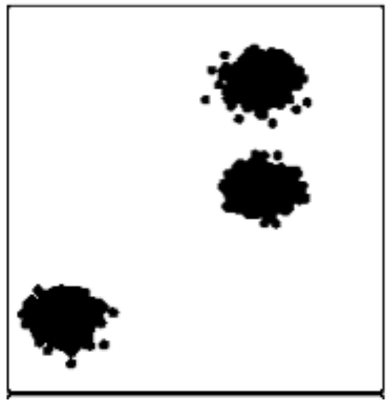
The Data Mining World — Clustering



From **scikit-learn.org**

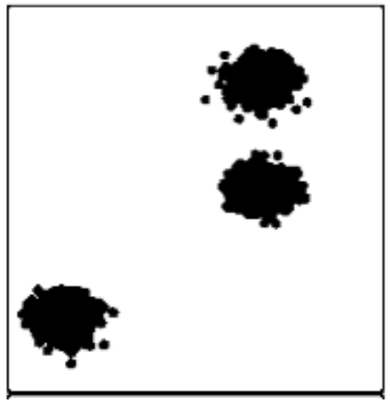
Clustering is an unsupervised learning technique

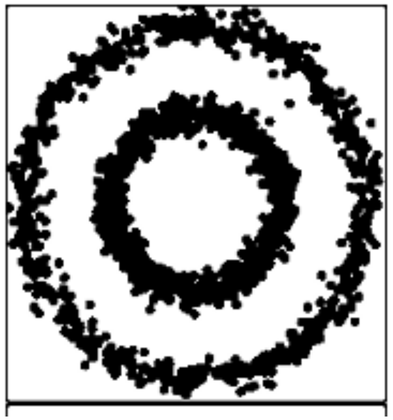
How many clusters are there?



Clustering is an unsupervised learning technique

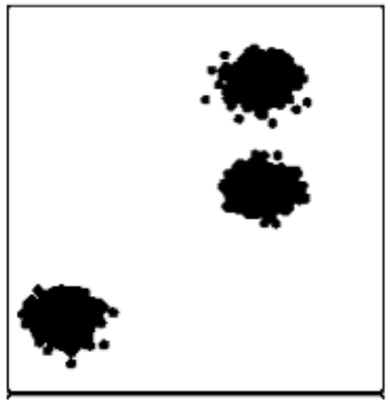
How many clusters are there?

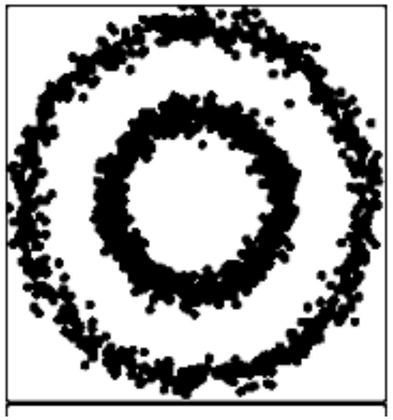




Clustering is an unsupervised learning technique

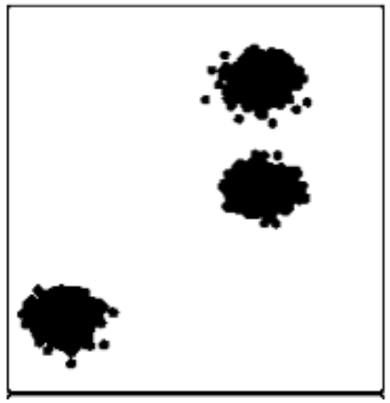
How many clusters are there?

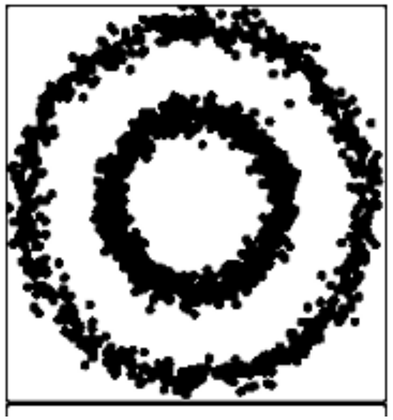


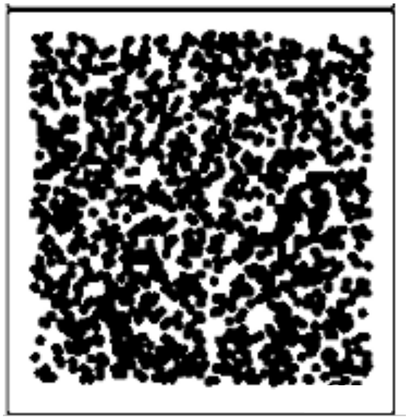


Clustering is an unsupervised learning technique

How many clusters are there?

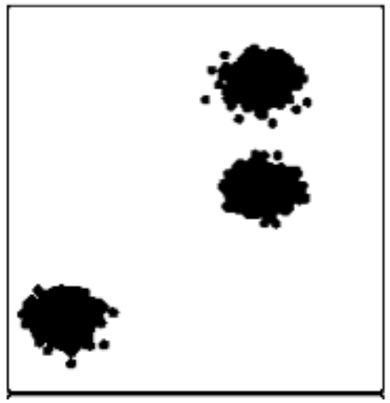
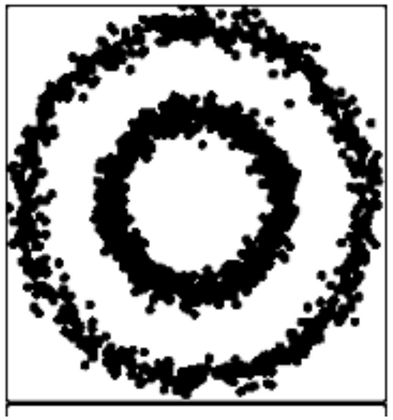


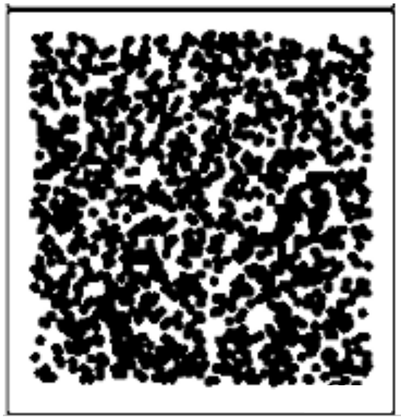




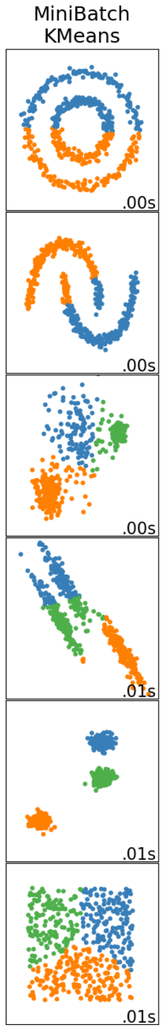
Clustering is an unsupervised learning technique

How many clusters are there?

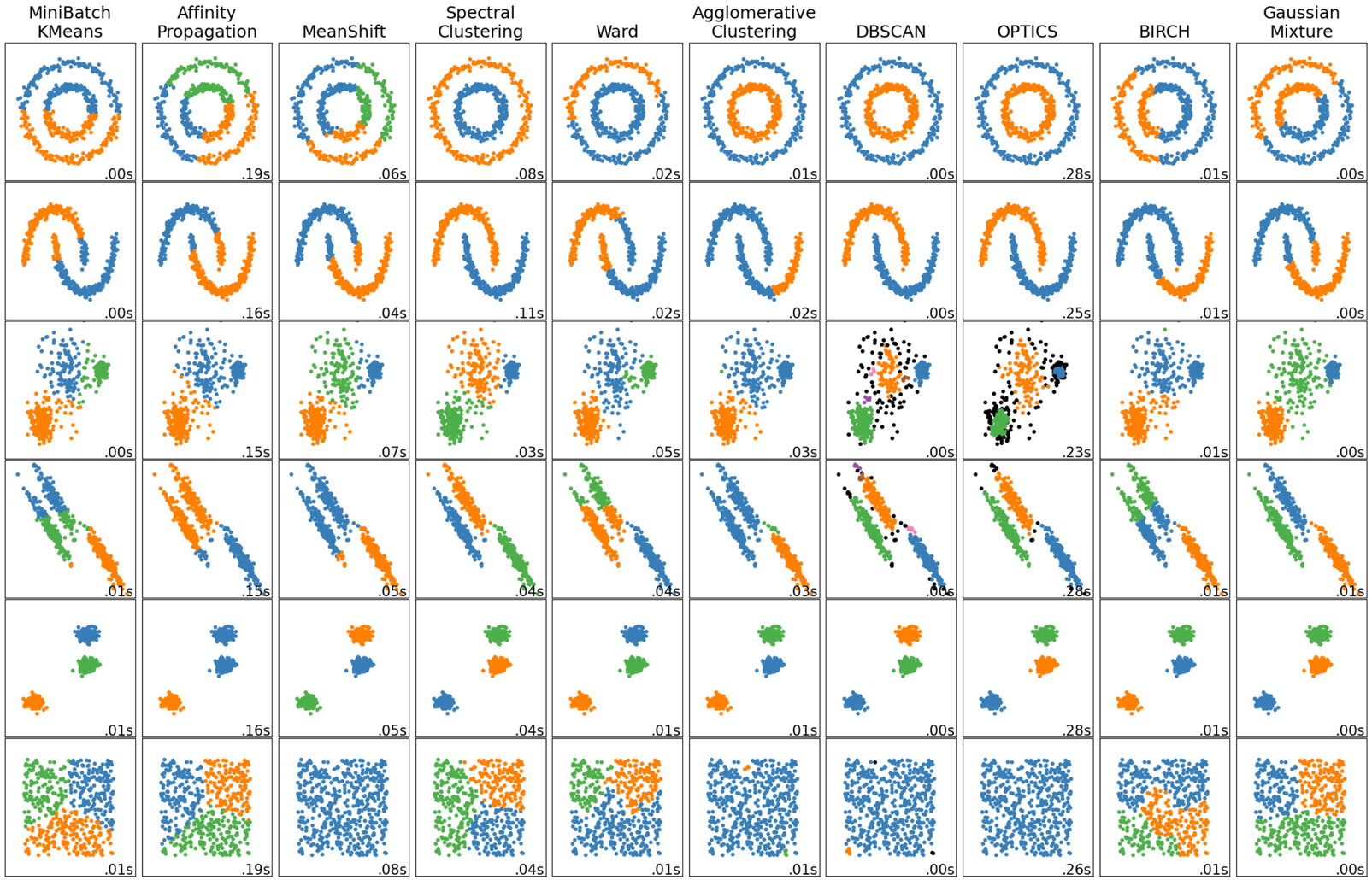






There are many different clustering algorithms

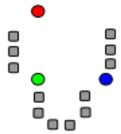


There are many different clustering algorithms

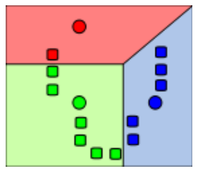


K-means, DBSCAN and spectral clustering

K-means

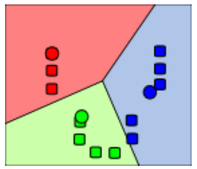


Initial guess



K-clusters are generated with the nearest mean

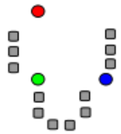
Centroid of the k-lcusters becomes the new mean



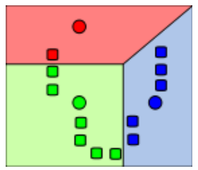
Iterate until convergence

K-means, DBSCAN and spectral clustering

K-means

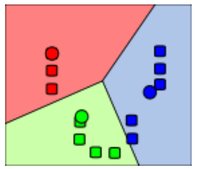


Initial guess



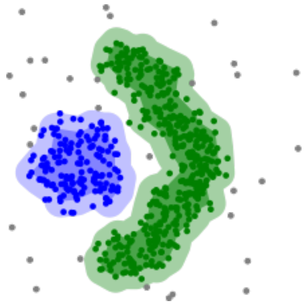
K-clusters are generated with the nearest mean

Centroid of the k-lcusters becomes the new mean



Iterate until convergence

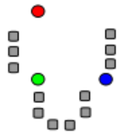
DBSCAN



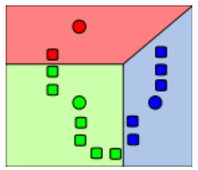
1. Find the points in the ε (eps) neighbourhood of every point, and identify the core points with more than minPts neighbours.
2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise assign it to noise.

K-means, DBSCAN and spectral clustering

K-means

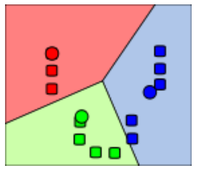


Initial guess



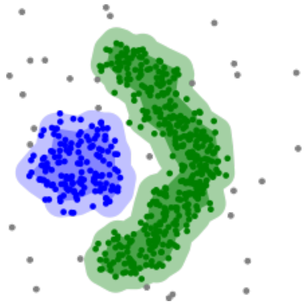
K-clusters are generated with the nearest mean

Centroid of the k-lcusters becomes the new mean



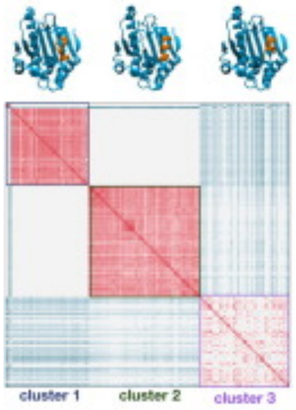
Iterate until convergence

DBSCAN



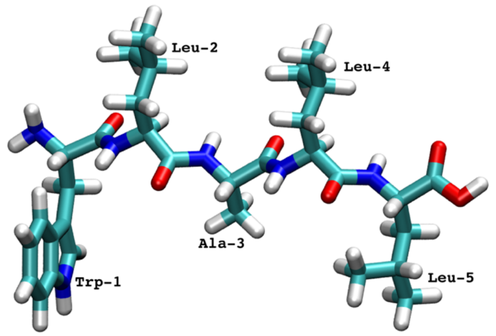
1. Find the points in the ε (eps) neighbourhood of every point, and identify the core points with more than minPts neighbours.
2. Find the connected components of core points on the neighbour graph, ignoring all non-core points.
3. Assign each non-core point to a nearby cluster if the cluster is an ε (eps) neighbour, otherwise assign it to noise.

Spectral clustering

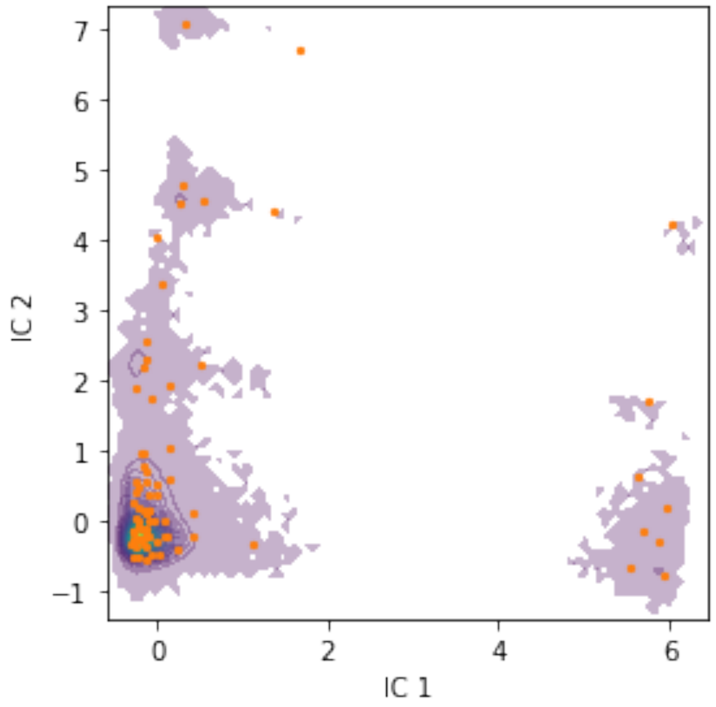


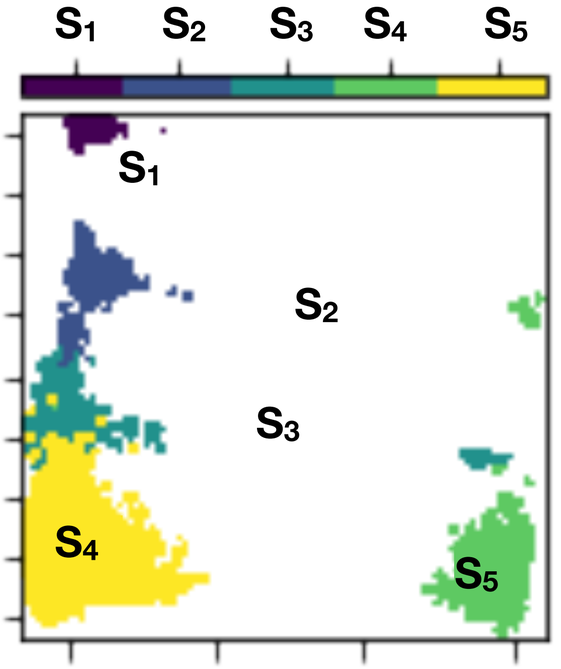
In spectral clustering clusters are found by doing an eigenvalue decomposition of the Laplacian

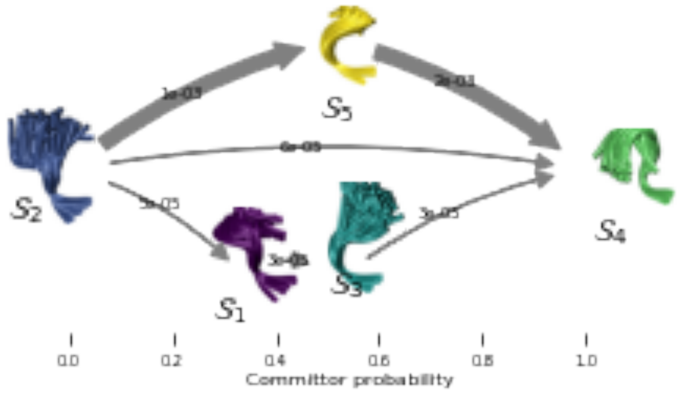
K-means example



Clustering is one of the first steps in building a Markov State Model







Post-its